DEVELOPING CONFIDENCE IN MEASURES OF EFFECTIVENESS FOR COMPLEX DEFENSIVE SYSTEMS
(WITH APPLICATION TO THE BALLISTIC MISSILE DEFENSE SYSTEM)

John F. Lyons

Johns Hopkins University
Applied Physics Laboratory
Laurel, MD 20723

Abstract

Defensive systems must be evaluated in the context of the threat. Complex threats and a range of scenarios require a computer model to estimate system effectiveness, often without use of probability distributions. It is important to evaluate system effectiveness in terms of statistical confidence to support informed decision-making, but there are two problems:

- Several types of non-homogenous data, each with separate probability distributions, must be combined to determine overall confidence for a single weapon system;
- A process to use confidence-derived measures of effectiveness for all the systems that comprise the Ballistic Missile Defense System, minimizing statistical error, must be developed.

Use of engineering judgment is unavoidable in using and combining test results to estimate confidence. Failure categories for several systems are analyzed to show the historical context that can aid in supporting engineering judgment to extrapolate available data to performance estimates of systems under development. Several means of eliciting and applying engineering judgment will be given.

Properly using probability distributions in Monte Carlo models is approached through the use of notional examples that illustrate the differences in output resulting from the number of runs, the type of statistical errors that can be made, and the effect of those errors.

Introduction

Defensive systems, unlike offensive systems, must be evaluated in the context of the threat that they are defending against. The threat may be complex and can vary over a range of scenarios, thus requiring a computer model to estimate system effectiveness or assess system capability. Complex defensive systems, such as, the Ballistic Missile Defense System (BMDS), use various models for evaluation, which must be used to provide statistical confidence, in performance estimates. It is important to evaluate system effectiveness in terms of uncertainty (confidence in the estimate) in order to support informed decision-making, but there are two problems:

- Top-level measures of effectiveness, such as, the probability of negating a threat, include multiple variables that are in themselves functions of time, accuracy, and reliability each with different probability distributions. Relating the uncertainty from test results on individual variables to establishing the confidence in a top-level measure is a complex, difficult process. The process can be reversed from evaluation requirements flowed down to determine the sample size needed to meet those requirements. A notional example applicable to BMDS will be developed.

- Several types of data: engineering, component, integration, simulation and system-level tests, each with separate probability distributions, must be combined with engineering judgment to determine overall uncertainty. This uncertainty must be expressed in terms of probability distributions to derive overall statistical confidence. Several examples using historical data will be shown.

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
29JUL2002	N/A	-

Title and Subtitle Developing Confidence in Measures of Effectiveness for Complex Defensive Systems (With Application to the Ballistic Missile Defense System)	**Contract Number**
	Grant Number
	Program Element Number
Author(s) Lyons, John F.	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Johns Hopkins University Applied Physics Laboratory Laurel, MD 20723	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es)	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)

Distribution/Availability Statement
Approved for public release, distribution unlimited

Supplementary Notes
See Also ADM201460. Papers from Unclassified Proceedings from the 11th Annual AIAA/MDA Technology Conference held 29 July - 2 August 2002 in Monterey, CA.

Abstract

Subject Terms

Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU

Number of Pages
7

Life Cycle Performance Estimation

Performance estimates of the expected deployed estimates can --- and must --- be done throughout the system life cycle. For the Missile Defense Agency (MDA), the elements of the BMDS, THAAD, PATRIOT, Ground Midcourse Defense system (GMDS), Airborne Laser (ABL), BMC3, et al, are in various states of the system life cycle from conceptual to operationally deployed. Any overall evaluation will depend on combining data from disparate sources with various levels of confidence in those estimates.

For a defensive system negating the threat will require a combination of the equipment working reliably, accurately, and in a timely manner. To estimate overall confidence will require probability distributions for all parameters: reliability, accuracy, and timelines. There is no one solution or approach to this problem. Though Bayesian techniques certainly permit estimating probability distributions at any level of performance, the basic approach proposed herein is to keep each uncertainty (confidence) estimate as close to being related to a direct test data source as possible.

In Part 1 of this paper, an example will be carried through to illustrate some of the issues involved. In Part 2 an example of how disparate data sources can be used in conjunction with expert judgment to provide additional confidence will be given.

Part 1: Using the Derived Distributions to Obtain an Overall Capability Estimate with Confidence

Systems can be developed to an Evaluation Requirement (one that contains a specified level of confidence) against a specified threat. A requirement for BMDS could take either of the forms below:

- **Given the necessary support and a firing doctrine of four interceptors per threat against a threat size of (S), the system shall provide protection for all 50 states with probability of negation of (X) with a confidence of (Y).**

- **Given the necessary inter-operating support and a firing doctrine of four interceptors per threat, the probability of negation for the total system shall be 0.X, known with a confidence of 0.Y.**

A difficulty with any requirements statement that is stated in the context of a scenario is that there are two sensible interpretations: the most demanding scenario case defines the measure of effectiveness (implied by the first type of requirements statement), or the average performance over all scenarios defines the measure of effectiveness (implied by the second type of requirements statement).

A second approach is to estimate performance capability if the system is called upon to perform (contingent deployment) before it is evaluated to a formal Operational Requirement. This type of approach could take the following form:

System Technical Objective

- **The initial goal is to develop an early capability that the estimated single shot reliability be known with X% confidence to be within 0.xx.**

This type of statement does not set a specific performance value to be met, as at the end of this phase of development is a system that has some initial capability, if needed, that will then transition to acquisition for development of the final system under a set of formal Operational Requirements. An example will clarify:

- The initial goal is to develop an early capability that the estimated single shot reliability be known with 90% confidence to be within 0.25.

The following table shows the success rate for the number of tests that would meet this criterion.

Demonstrated Results	4/8 = 0.50	6/10 = 0.60	7/10 = 0.70	8/10 = 0.80	8/8 = 1.00
Lower 90% Confidence	0.25	0.35	0.45	0.55	0.75

The number of tests varies from 8-10 based on the results; results lower than 0.5 are certainly possible, but would inevitably result in a major corrective action that would change the nature of the system and require a new basis for an estimate. The key to planning is knowing the confidence in the estimate, more than the actual estimate itself. If expected performance is known with high confidence, then the firing doctrine can be planned with more confidence.

A simplified example will show how first, to derive the overall distribution for the probability of negation and then, flow down from the above requirements statement to determine a set of test sizes for the various parameters.

First, assume a single threat against a single target.

- **Threat** Ballistic missile SADDAM
  - NBC payload
  - 10 balloon radar decoys, one light-weight replica
  - Threat range 9970 km (Baghdad to Washington D.C.)
  - Time of flight 1881 sec (31 min)
  - Velocity at burnout: 7.5 km/sec

To build a model, a functional description of the system is needed and to run the model a scenario is required. Note that in order to develop confidence limits all input parameters must be in terms of a probability distribution.

- **Defensive System**
  - Stationed 5400 km from Washington D.C.
  - Interceptor velocity: 6.0 km/sec - Normal distribution (6.0, 0.03)
  - Radar threshold detection: 0 dBSM @ 7000km - Normal distribution (7000, 500)
  - Balloon: discrimination probability 0.9 - (K-factor ± 0.2)
  - Light-weight replica: discrimination probability 0.5 - (K-factor ± 0.5)
  - Reliability: 0.70 - Binomial distribution (7 successes in 10 attempts)
  - Accuracy: Bivariate normal distribution (7 tests)

Each one of the above distributions could have been derived based on direct test data, but the data would only apply to the conditions under which the tests were conducted. Extrapolating to other conditions would require use of engineering models that could be extremely complex.

Prior to running a simulation to determine the overall distribution for the probability of negation, a sensitivity analysis could be done to determine which parameters had little to no effect on the results. For this simplified scenario, the following sensitivity results would apply to each of the distributions as follows:

- Interceptor velocity: No effect; few seconds variation plus or minus, with a hundred seconds of variation in intercept time available.
- Radar detection threshold: Small effect; worst-case radar detection, in this case, would occur well within time to permit interceptor engagement. (There is no possible second launch based on observing the first interceptors results. The launch site is 5400 km from the minimum interceptor point allowing no time for a second salvo.)
- Balloon discrimination: Large effect; amplifies the effect of interceptor reliability by making intercepts of the target object less likely.
- Lightweight replica discrimination: Large effect; amplifies the effect of interceptor reliability; 0.5 probability that the first interceptor would engage the decoy rather than the target object.
- Interceptor reliability: Moderate effect; if discrimination were not necessary, a one-sigma variation in single shot interceptor reliability would cause a change in overall probability of negation (using four interceptors) from 0.961 to 0.999.
- Accuracy: No effect; This is a hit-to-kill interceptor; any hit within four CEP at the closing velocity of 13.5 km/sec, for this geometry, would destroy any NBC payload

Effectiveness Calculation

The next step is to run a simulation that can combine all the individual distributions to derive the overall distribution on the desired parameter: probability of negating the threat. For ICBM trajectories, a fairly detailed 6 degree-of-freedom simulation would be necessary to account for earth rotation and determine if an intercept were possible. For this calculation with the given parameters, detection would occur well within sufficient time making the only parameters that have to be considered, the uncertainty in interceptor reliability and discrimination capability.

The table below shows that the lower 90% confidence limit based on ten tests for five through 10 successes. Assume there were 3 successes in five tests, with the prior assumption that the final deployed reliability would be uniform above 0.60 (Part 2 illustrates this example); this would amount to a combined sample of approximately ten tests with an estimate of 0.70 and a lower 90% confidence limit of 0.45. With a single shot reliability of 0.70 there would be a $[1-(1-0.70)^3]$ probability of at least one of the four interceptors negating the threat. The lower 90% confidence limit that one in four interceptors will be successful is 0.93.

Combining a demonstrated test result of 3 successes in five attempts, with a judgment based on historical data that all deployed systems achieved reliability above 0.6, the 90% lower confidence that the system can negate the threat is 0.93. This could serve as an example of projecting expected deployed performance based on combining early test results with historical data tests near the beginning of the system life cycle.

Single shot Reliability	0.5	0.6	0.7	0.8	0.9	1.0
Lower 90% confidence	0.26	0.35	0.45	0.55	0.66	0.79
Prob. At least one hit in 3 shots	0.87	0.94	0.97	0.99	1.00	1.00
Lower 90% confidence	0.63	0.71	0.75	0.80	0.80	0.80

With a firing doctrine of three interceptors per incoming threat, the confidence that at least one of the three interceptors will destroy the target is significantly higher than the confidence for a single interceptor. The overall confidence is a function of the combined confidence in all the parameters that affect the outcome as given by the functional decomposition and requires the use of a detailed simulation. To show the effect of adding parameters, the probability of correct discrimination, which can be calculated without resorting to a simulation, will be included.

With a firing doctrine of three interceptors, the above table shows that the probability of at least one interceptor destroying the target, if the interceptor reliability is 0.7, is 0.973. The effect of the balloon decoys with a probability of correct discrimination of 0.9 would be to lower the interceptor reliability to 0.9 X 0.7 = 0.63. (This assumes that destroying one balloon of ten would not change the probability of discrimination). This would lower the probability of negation by at least one of three interceptors negating a single threat to 0.950.

The effect of a lightweight replica, with a probability of discrimination of 0.5, is more complicated. Assuming the first reliable interceptor correctly discriminates the threat object and the replica from the balloons, there is a 0.5 probability that it will destroy the replica. The second reliable interceptor will destroy the object not destroyed by the first. The effect of all possible sequences of four interceptors would be to lower the probability that at least one interceptor of four will destroy the threat to 0.880.

Effect on test sizing

The above process can be reversed: starting with the requirement and then determine the number of tests required to meet the desired confidence. There is no easy answer to combine the test uncertainty of multiple variables --- and for some variables: discrimination, aimpoint selection, lethality are themselves functions of multiple variables --- that can totally dominate the answer. Narrowing down the effects of test uncertainty to a useful range requires an extensive sensitivity analysis, a simulation that combines probability distributions in terms of the desired top-level measures, and delimiting the effect of uncertainty of certain dominating variables by isolating the effects of those variables.

For BMDS, there is no direct test possible of all the factors comprising the overall probability of negation, which must be derived in the context of a scenario using models and simulations; however, the one variable that is independent of the scenario is reliability of those items necessary to make an intercept. This would include a sensor to detect the threat, communications to the firing unit, and the interceptor. The sensor and if necessary communications equipment operate continuously and, therefore, can be directly, extensively, and non-destructively tested in an operational mode. Interceptor in-flight reliability is the key parameter that, for direct testing, must be destructive. Another critical parameter, but not a reliability factor, is the probability of correct discrimination can be tested during flight tests, however, there are a broad spectrum of conditions that will not be tested and can only be tested through the use of high fidelity or hardware/software-in-the-loop models. Models that are based on fundamental physics can add confidence when calibrated at a few points to verify the physical representation. Models of this type can partially substitute for testing by adding predictive confidence.

Historically for most systems, more reliability failures occur early in the system life cycle and it is unlikely that the above success rates would be demonstrated before the system is fully deployed. This means that system level test data will have to be combined with other data, as described in the second part of this paper, to increase confidence with fewer tests until a larger number of tests are available. (Of course, the system may not achieve a better than 0.684 probability of success in its life cycle, but history does show that most systems that are deployed do so.

The next step is to be able to use probability distributions that are a measure of uncertainty for all of the parameters, in a simulation that can produce the desired measure of effectiveness, in our example the probability of negation. JHU/APL has developed a simulation expressly for this purpose that is capable of incorporating updated test results with prior distributions.

Part 2: Use of Historical Test Data

Elicitation of expert judgment requires close collaboration between statisticians and subject matter experts. In this paper, examples of eliciting expert judgment will be to estimate deployed system effectiveness for the ground midcourse defense system (GMDS) part of the Ballistic Missile Defense System (BMDS) to defend against ballistic missiles launched toward the U.S.

The application of engineering judgment is unavoidable in using and combining disparate test results to estimate uncertainty. The first step is to ensure that disparate test results are place in failure categories that are reasonably homogeneous: design failures, random manufacturing/production failures, and non-random (a failure that applies to only a known unit or portion of the force, aging would be an example of a non-random failure), as shown in Table 1. The categories of failures were derived from analysis of 12 offensive and defensive missile systems that were analyzed to show the historical context that can aid in supporting engineering judgment. This approach can be used to combine available test data to derive system performance estimates in any stage of the system life cycle.

Dealing with the Uncertainty of Engineering Judgment

Confidence limits are a measure of uncertainty in a statistical estimate that may come from several sources: observational error, statistical inferences, cognitive and institutional bias, and the number of applicable tests. If the desired measure of effectiveness can be determined from direct test results, then the number of tests to meet stated confidence limits can be readily derived. If direct test results of a high-level measure of effectiveness are not possible, as is the case with the BMDS, then the number of tests must be indirectly inferred from combining data from multiple sources. In the absence of direct test results, engineering judgment, in some form, either to estimate the degree of applicability of subsystem tests to the system level, the applicability of similar system data, or simply estimating the uncertainty in statistical terms, will be required. In any case, the amount of uncertainty will have to be translated into the parameters of a probability distribution to be of use.

Table 1
Flight Failures by Category and Year in the Life Cycle
(Based on 12 Weapon Systems –
6 Defensive, 6 Offensive)

Year	Design Failures	Perform-ance Limit Failures	Random Failures	Non-random Failures
1	2	4	22	3
2	10	1	18	1
3	11	3	23	0
4	6	1	16	1
5	2	6	26	1
6	0	2	20	1
7	1	2	17	1
8	0	4	19	5
9	0	4	23	3
10	0	2	30	2
11	0	2	23	3
12	0	1	21	0
13	0	3	23	1
14	0	5	23	0
15	0	6	28	0
16	0	3	19	1
17	0	1	25	0
18	2	3	18	0
19	0	4	20	0
20	0	5	13	0
21	0	0	12	0
22	0	0	14	0
23	0	0	10	0
24	0	3	8	0
25	0	2	12	0
26	0	1	9	0
27	0	0	6	0

The following examples illustrate several means of translating engineering judgment, component data, or similar system test data into probability distributions.

- **Example 1:** Eliciting expert judgment on the probability of a propulsion failure for a new generation defensive missile.
    - The expert has data from ground and sea-based defensive missiles that shows 60 propulsion failures in 5000 tests (0.988) that should be applicable. The statistician would ask, what is the expected failure rate and how confident is that judgment. The expert's answer if all the tests were judged applicable would be translated to: a confidence of 95%

that the probability of a propulsion failure is no worse (upper limit) than 0.05.

This example will be used to discuss several aspects of soliciting expert judgment. First, it is unlikely that an expert would be comfortable stating the answer in strictly or directly in statistical terms or that all the data would be judged directly applicable. More likely would be a vague statement: "There have been a lot of tests with very few failures in similar systems." There are several modes of response that can be used to assist in eliciting expert judgment:

- Likert scales: On a scale from 1 to 10 this is a 9
- Odds ratios: I would expect 1 failure but no more than 5 in a hundred.
- Comparative: No worse than system A and that was extremely high.
- Fuzzy: Expect propulsion probability to be something above (or around) 0.98.
- Probability: 0.99 ± 0.01

Eliciting expert judgment is an iterative process and the amount of uncertainty differs for each of the above responses, but each response would have to be translated to a statistical confidence statement of the type shown in the example in order to express the uncertainty in terms of an equivalent test sample size. Care must be taken not to elicit information that is based on more than is known that leads to a larger sample size than is warranted. Mathematically, this can be done through using methods that lead to maximum entropy (least amount of information) distributions.

The above distribution would be binomial and an equivalent (pseudo) sample size can be determined by knowing any two of the three parameters: lower confidence limit at X%, upper confidence limit at Y%, or the estimated mean. In the example, the upper 95% confidence limit is 0.05 and the mean failure estimate is 60/5000 = 0.012. This leads to a sample size of 99, which appears reasonable in view of the large database. This calculation should be made known to the expert so that they know the implications of their statements, especially in the case of having thought they were more uncertain than their answer implied.

- **Example 2:** Eliciting expert judgment on the overall probability that would be expected after the system has been deployed for a period of time.
  - The data from Table 1 shows that:
    - Design failures occur initially and are eliminated
    - Performance limit failures occur at a low rate
    - Random failures occur at a rate from 0.10-0.20 decreasing over time.
    - Non-random failures are rare (<0.01).

In this case, if a statistician is doing the elicitation, there may be several methods to examine the data to assist the expert. First, the database has been partitioned to obtain as homogeneous sample as possible. Differences among systems can be statistically analyzed (hierarchical analysis), and then, through a regression analysis, a statistical model developed. The model can be partially validated by iteratively sampling part of the database, using the model to predict, and then comparing the prediction to the known results. After reviewing the data, the expert's judgment may be: reasonably confident it is no worse than 0.6 and probably around 0.8 or slightly above. Fuzzy logic can be used to convert this to a probability distribution; 75% confidence that the lower limit is 0.6 with an estimate of overall effectiveness of 0.8. This translates to a sample size of 7 and would increase rapidly if initial test results confirmed this.

- **Example 3:** Again, eliciting an expert judgment on what is the expected deployed performance capability.

  Historical data shows that defensive missile systems of all types were at least 0.6 in overall system effectiveness when deployed. Further, historical data shows that once deployed performance increases.

This example is meant to show that even if a uniform distribution (maximum uncertainty) is used, this provides some information when used in a Bayesian formulation. If the assumption was a uniform distribution from 0.0 to 1.0, the full possible range of probability; this adds an equivalent sample size of one. Thus, if system tests resulted in 3 straight successes, the Bayesian reliability estimate, assuming a prior uniform distribution from 0.0 to 1.0, would be ¾ = 0.75 and <u>not,</u> 3/3 = 1.0 as would be the case for standard statistics. The above assumption could be described by a distribution with some low probability (5%) uniform from 0.0 – 0.6 and the remainder (95%) uniform from 0.6 – 1.0 leading to a pseudo sample size of approximately 4. Depending if test results that confirmed the prior estimate (>0.6) the confidence would increase more rapidly and if test results varied from the prior estimate, more slowly.

Summary

This paper has presented techniques for developing confidence based, deployed performance estimates using early test data combined with engineering judgment for a complex system of systems. The approach used is to partition all available test results into homogeneous subsets, use historical data for similar systems in a limited manner, and apply Bayesian techniques to develop confidence in the performance estimates. As more test results become available over the system life cycle, the more heavily the test results are weighted until deployed test data is available when confidence will be derived solely from test data.